

Chapter 2

Sampling Survey of Heavy Metal in Soil Using SSSI

Aihua Ma, Jinfeng Wang, and Keli Zhang

Abstract Much attention has been given to sampling design, and the sampling method chosen, directly affects sampling accuracy. The development of spatial sampling theory has lead to the recognition of the importance of taking spatial dependency into account when sampling. This research study uses the new Sandwich Spatial Sampling and Inference (SSSI) software as a tool to compare the relative error, coefficient of variation (CV), and design effect of five sampling models: simple random sampling, stratified sampling, spatial random sampling, spatial stratified sampling, and sandwich spatial sampling. The five models are each simulated 1,000 times, with a range of sample sizes from 10 to 80. SSSI includes six models in all, but systematic sampling is not used in this study, because the sample positions are fixed. The dataset consists of 84 points measuring soil heavy metal content in Shanxi Province, China. The whole area is stratified into four layers by soil type, hierarchical cluster and geochronology, and three layers by geological surface. The research shows that the accuracy of spatial simple random sampling and spatial stratified sampling is better than simple random sampling and stratified sampling because the soil content is spatially continuous, and stratified models are more efficient than non-stratified models. Stratification by soil type yields higher accuracy than by geochronology in the case of smaller sample sizes, but lower accuracy in larger sample sizes. Based on spatial stratified sampling, sandwich

A.H. Ma (✉) • K.L. Zhang

School of Geography and Remote Sensing Sciences, Beijing Normal University, 100875
Beijing, China

e-mail: maaihua23@126.com; keli@bnu.edu.cn

J.F. Wang

Institute of Geographic Sciences and Nature Resources Research, Chinese Academy of Sciences,
100101 Beijing, China

e-mail: wangjf@lreis.ac.cn

sampling develops a report layer composed of the user's final report units, allowing the user to obtain the mean and variance of each report unit with high accuracy. In the case of soil sampling, SSSI is a useful tool for evaluating the accuracy of different sampling techniques.

2.1 Introduction

Much progress has been made in developing tools to assess estimation accuracy and the relevance of chosen models, such as spatial autocorrelation precision (Cressie 1991; Wang et al. 2002; Haining 2003; Li et al. 2005) or error variance (Wang et al. 2010). Uncertainty in a spatial sampling survey can be reduced with high-quality information, improvement of statistical estimation accuracy, and increasingly precise geo-statistical software. The choice of a relevant sampling model, however, is essential to the quality of the analysis and that which is learned from the study.

Many different sampling methods exist, and selecting one, relevant to a specific problem, is always an important choice for the scientist, who wants to have the most valid statistical inference analysis possible. The user can decide to study the problem with, for instance, a simple random sampling model. This involves selecting a fixed number of units from a population, each with an equal chance, or the selection of a stratified sampling model to assess the relationship between samples and a stratified information layer (Cochran 1977). The method choice really depends on the problem, on the objective of the study, or on the available dataset. Having methods with which to assess the fit of the sample dataset with the sampling model would be useful to the scientist, and in all likelihood improve the statistical analysis and refine results.

In this study, two approaches for matching sampling effort to accuracy are chosen: a classical approach, which ignores spatial dependence between observations; and a geo-statistical approach, which accounts for spatial dependence (Hedger et al. 2001). Accounting for spatial dependency decreases the standard error when the environmental variable is spatially correlated (Wang et al. 2002; Haining 1988).

Sandwich Spatial Sampling and Inference (SSSI) is a professional spatial sampling and statistical inference tool. It can be used for environmental, resource, land, ecological, social, and economic sciences and practice sampling design and statistical inference. The objective of this study is to evaluate the accuracy of different sampling methods by comparing several indicators, such as relative error, coefficient of variation (CV), and design effect, using SSSI. Because the sampling model and number of samples (or sample size) are the two most important factors determining sampling efficiency (Chen et al. 2009), this study compares four models with a range of sample sizes from 10 to 80. The dataset is a survey of heavy metal content in soil in Shanxi province, China.

2.2 Material and Methods

2.2.1 Data and Site Description

The investigation and environmental sampling was carried out in 2002 and 2003. Two counties, Zhongyang and Jiaokou, in Shanxi Province were selected as the research area. There are seven administrative villages in each county and each has a high incidence of birth defects. Samples were collected using administration villages as units. Sample sites, consistent with birth defect epidemiological survey sites, were considered, so there exist some gaps within the data set, such as no samples in the western region. Soil samples were collected far from soil disturbance plots, such as those affected by residences and roads, to ensure, as far as possible, the true representation of local environmental characteristics. Samples were collected in most villages, resulting in a total of 84 points in all. Well-mixed cultivated land surface soil samples were acquired with a shovel 10–20 cm deep. The samples were dried, cleaned of organic debris and impurities, ground with a mortar, sifted with a 100 mesh aperture, and weighed at about 0.1 g soil 3 copies. They were then mixed with 3 ml of Glass Acid, 1 ml perchloric acid, and 1 ml hydrofluoric acid and placed into an oven, for 5 h, to bake and nitrify. Finally, they were put on an electric heating board for 1–2 h, mixed with water to give a uniform volume of 10 ml, for testing (Zhang et al. 2008). An ULTIMA inductively coupled plasma spectrometer was used to measure 16 elements in the soil: Al, As, Ca, Cu, Fe, K, Mg, Mo, Na, Ni, Pb, Se, Sn, Sr, V, and Zn. In this paper, Mo contents are used as sampling data.

Heavy metal content in the soil primarily depends on the natural levels of soil-forming parent materials and the absence of external contaminants (Wang et al. 1992). Through the process of weathering and soil formation, the concentration of these elements will change, but the relationship between parent material and their concentration in the soil is still relatively close. There are five main soil types in pedogenesis: cinnamon, chestnut cinnamon, loessial soil, regosols soil and Chao soil. In this way, the area is stratified by soil type and geological surface. It can also be stratified geochronologically, because elements can be introduced at different times. There are eight layers, but in order to keep the sample size reasonable in each layer, the region was stratified into four layers. The fourth stratification method is hierarchical clustering accomplished with SPSS.

2.2.2 Sampling Method and Algorithm

Five models in SSSI were used: simple random sampling, stratified sampling, spatial random sampling, spatial stratified sampling, and sandwich spatial sampling. The first two models took the classical approach, and the last ones took the geo-statistical approach. The use of different sampling designs and methods may

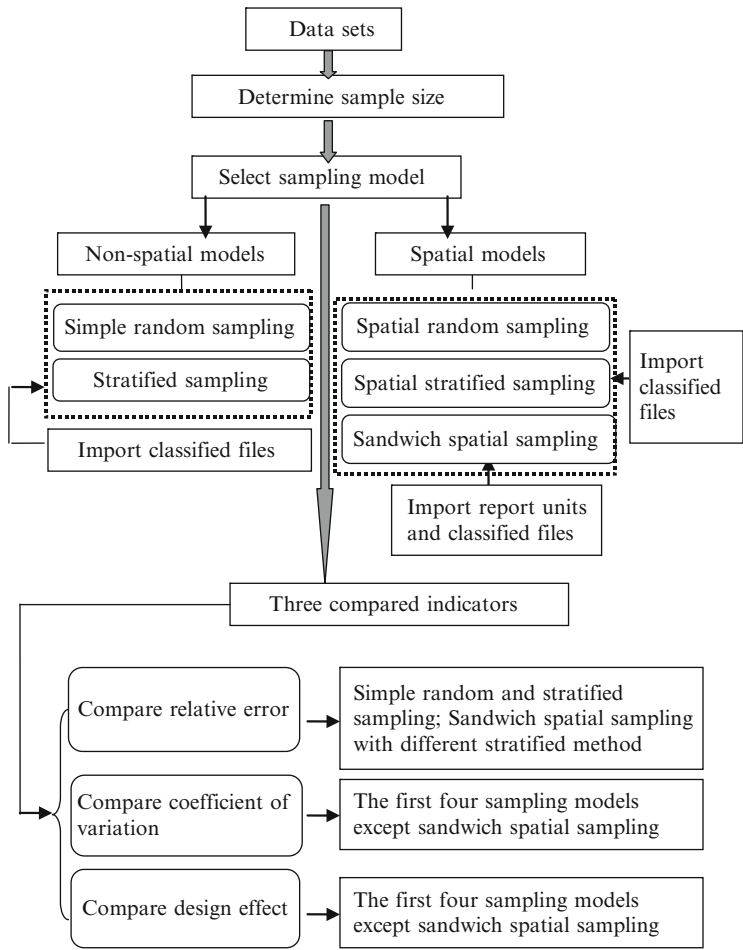


Fig. 2.1 A flow chart of the steps of this study

affect the accuracy of the analysis of samples collected in the same area (de Zorzi et al. 2008). In these cases, three indicators were selected to compare the efficiency of different sampling models. To select the three indicators, the models were simulated 1,000 times, each with a range of sample sizes from 10 to 80. To ease reader’s understanding, Fig. 2.1 gives a flow chart of the steps taken:

2.2.2.1 Relative Error

Generally, the accuracy of sampling is higher with larger sample sizes, and the sample mean is nearer to the population mean. But at a certain sample size, the sampling method can affect the accuracy, so relative error (d_r) is a good indicator to

judge which sampling model is more accurate. Relative error compares the difference between the sample mean and its true mean, so the estimated relative error is defined as:

$$d_r = \frac{abs(\bar{Y} - \bar{y})}{\bar{y}} \tag{2.1}$$

where $\bar{y}$ = sample mean
 $\bar{Y}$ = observable population mean.

The former is estimated by the different models and the latter is acquired by counting the entire observable metal contents.

Figure 2.2 compares d_r of the simple random sampling and of the stratified sampling. The sandwich spatial sampling model is the newest method in the SSSI software. It has the same accuracy as spatial stratified sampling, but it considers the report layers, which can be in any unit, for example, a county border, provincial boundary, watershed, or artificial grid (Wang et al. 2009). Table 2.1 shows the values of the whole relative error of report layers.

2.2.2.2 Coefficient of Variation

The coefficient of variation is a statistical measure of the dispersion of data points in a data series around the mean. It is calculated as follows:

$$\text{coefficient of deviation} = \frac{\text{Standard deviation}}{m} \tag{2.2}$$

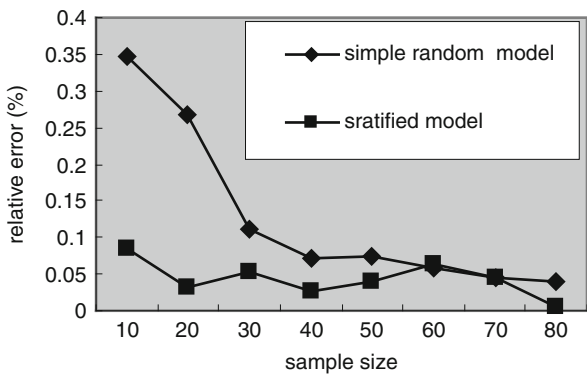


Fig. 2.2 Comparison of the relative error of simple random sampling and of stratified sampling

Table 2.1 Relative error of two kinds of report layer

Report layer	Stratified by soil type	Stratified by geochronology
Administrative villages	0.180	0.085
Grids	0.066	0.052

$$\text{Standard deviation} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

(2.3)

where m = sample mean
 x_i = value of element content
 n = sample size.

In this paper the dispersion of each estimator from its true value during the 1,000 simulations for each sampling method and sample size is compared (Chen et al. 2009).

Figures 2.3, 2.4, 2.5 and 2.6 show the CV from four sampling models of Mo elements, stratified by soil type, geological surface, hierarchical cluster, and geo-chronology. Figure 2.7 shows which stratified method is the most efficient.

Fig. 2.3 Mo (stratified by soil type)

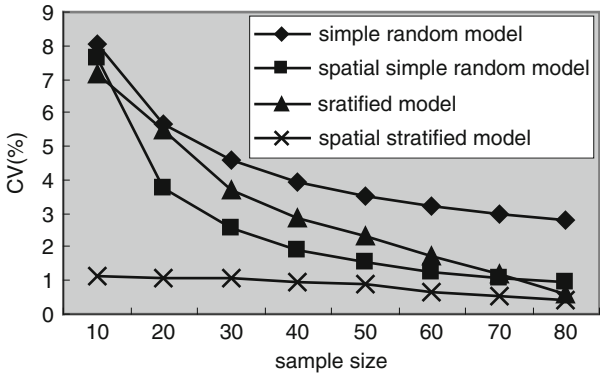


Fig. 2.4 Mo (stratified by geological surface)

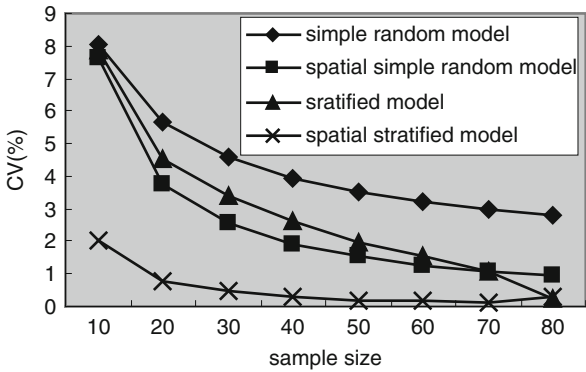


Fig. 2.5 Mo (stratified by Hierarchical cluster)

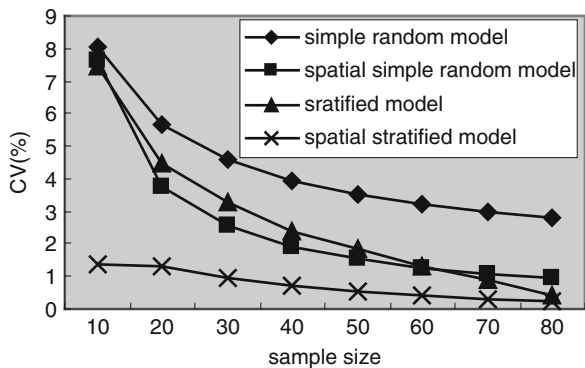


Fig. 2.6 Mo (stratified by geochronology)

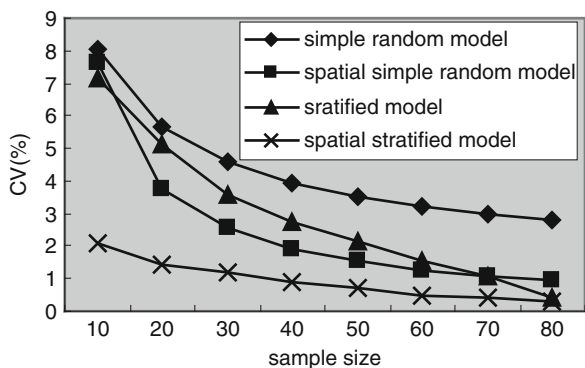
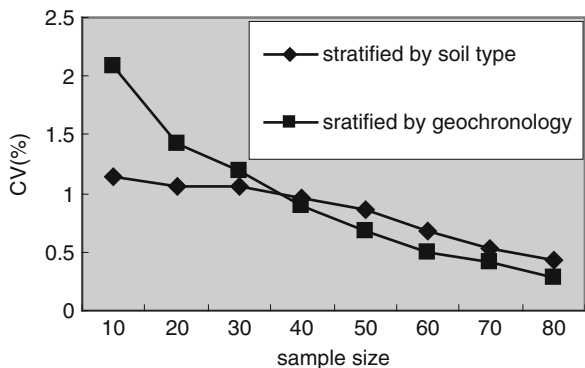


Fig. 2.7 Mo compare stratified method soil type and geochronology



2.2.2.3 Design Effect

The design effect is the ratio of variance of the estimate obtained from the (more complex) sample to the variance of the estimate obtained from a simple random sample of the same number of units (Cochran 1977). This indicator also enables the comparison of the efficiency of different sampling models.

2.3 Results and Discussion

Samples are spatially distributed, so geo-statistical approach which takes spatial dependency into account is used besides classical approach. Finally, five models of SSSI software have been chosen for comparison, while not all of them is used to compare the above three indicators. Spatial random sampling is the same as simple random sampling as regards evaluating the sampling mean, and spatial stratified sampling is also the same as stratified sampling. Hence only the relative error of the simple random sampling and that of the stratified sampling is compared. When importing report units in sandwich spatial sampling, each report unit has individual relative errors, thus all relative errors in the report layers are compared. Sandwich spatial sampling has the same accuracy as the spatial stratified sampling in coefficients of variation and design effect, hence these last two indicators are compared in four sampling models, only.

2.3.1 Analyses of the Relative Error

The simple random sampling model is a method in which each sample has an equal chance of being selected. Stratified sampling divides the whole area into groups based on a special factor. This factor affects the distribution of the research object; there is little variation inside the layer but great variation between layers. Figure 2.2 shows that as the sample size increases, the relative error in the simple random sampling model, decreases. With smaller sample sizes, the simple random sampling model has lower accuracy than that of the stratified sampling model, and the interval is large. When the area is stratified by soil type, each sample is represented in every layer, although sample sizes are small. With larger sample sizes, the stratified sampling model fluctuates within a certain range, but is more accurate than the simple sampling random model from 10 to 80. Therefore, the stratified sampling model has a higher sampling efficiency than the simple random sampling model and results in a more representative sample.

Table 2.1 presents two kinds of report units: administrative villages and grids. Under normal circumstances, the layer created with the stratified sampling model is called the knowledge layer and the number of knowledge layers is less than the number of report layers. The mean and variation of each report layer are obtained with SSSI, and using Formula (2.1), each report layer's relative error can be calculated. The whole relative error can be found by combining layers. The values of the table are less than 10%, except for administrative villages, stratified by soil type.

2.3.2 Analyses of the Coefficient of Variance

Figures 2.3, 2.4, 2.5 and 2.6 show the coefficient of variance (CV) of Mo for the simple random sampling model, spatial random sampling model, stratified sampling model, and spatial stratified sampling model. CV is the ratio of spatial variation to the sample mean. For small sample sizes, the estimated sample mean is less accurate than that for large sample sizes. The CV decreases with an increase in the sample size for all sampling models. But the curve fluctuation ranges are different for each sampling model. The spatial stratified sampling model has the highest precision, indicated by its curve, which is below those of the other graphs, and the simple random sampling model has the lowest precision. The spatial random sampling model has higher precision at the middle sample sizes at about from 15 to 70. It can be concluded that the heavy metal contents in soil have spatial autocorrelation and that the precision of the different sampling models is impacted more by the spatial dependency than by the areal heterogeneity for most sample sizes.

Taking spatial dependency into account, the nearer the samples are to one another, the more similar their attributes and the greater their spatial correlation. Spatial mean variation is calculated from the samples with the following formula:

$$\text{Var}\{\bar{Z} - Z(A)/Z\} = \frac{1}{n} [\sigma^2 - E\{C(X, Y)\}] = \frac{\sigma^2}{n} - \frac{E\{C(X, Y)\}}{n} \quad (2.4)$$

where σ^2 = population variation

X, Y = random variables subject to uniform distribution in the area A

$C(X, Y)$ = the covariance of random variables X and Y .

Classical sampling mean variance is σ^2/n , and spatial sampling mean variance is the above formula with a decrease of $E\{C(X, Y)\}/n$ from classical sampling (Wang et al. 2009).

Spatial stratified sampling models take area heterogeneity and spatial dependency into account, not only to keep samples representative, but also to calculate sample autocorrelation. The sampling mean variance is smaller than any other sampling models, but the simple random sampling model does not account for area heterogeneity and spatial dependency, so the efficiency is lower. The stratified sampling and spatial random sampling models account for only area heterogeneity or spatial dependency, hence their efficiency is between spatial stratified sampling models and simple random sampling models. For small sample sizes, the stratified sampling model is better than the spatial random sampling model because it shows that the dependence of samples is not strong and the samples are more representative. With greater the sample size, the stronger the dependence. Although samples are more representative with the stratified sampling model, the impact degree is not larger than spatial dependence. When the sample size is about 60–70 (almost the population size) the spatial dependence reaches its maximum. It tends towards a certain level value, so there are some limitations for spatial dependence at small and large sample sizes.

As shown in Figs. 2.3, 2.4, 2.5 and 2.6, the accuracy of the simple random and spatial random sampling models is the same. A difference is shown in the stratified sampling model and spatial stratified sampling model because of the different stratification methods used. Selecting the best stratification method is important in the estimation of sampling precision. If the stratification is poor, the accuracy could be lower than that produced by simple random sampling. Figure 2.7 gives a comparison of the CV of two stratification methods. In the graph, stratification by soil type yields higher accuracy in the case of smaller sample sizes, than by geochronology, but lower accuracy in larger sample sizes. Knowledge layer is not necessarily spatially continuous, so samples are also not necessarily near each other even in the same layer. But the stratification of soil type is much more continuous than that of the geochronology type. We know that the closer the objects, the higher the similar degree of objects. When sampling a certain samples in each layer and the sample size is very small, a sample which obtains through stratification of soil type can be more representative, and it brings high sampling efficiency. In the larger sample sizes, however, geochronological stratification is better, possibly because the Mo content has been more strongly affected by geochronology. This effect is manifested,when the sample size reaches a certain degree.

2.3.3 Analyses of the Design Effect

Tables 2.2 and 2.3 show the ratio of mean variance for different sampling models to the mean variance from the simple random sampling model. The values in the tables are all less than one, indicating that the simple random sampling model (Srs) is the least efficient and the spatial stratified sampling model (SStrs) is the most efficient. Spatial autocorrelation slightly impacted the sampling accuracy.

The spatial stratified sampling model produces accuracy, higher than any other sampling models. In traditional stratified sampling method, objects within the same

Table 2.2 Design effect of spatial simple sampling, (spatial) stratified sampling by soil type and geological surface

Sample sizes	Models				
	Srs	StrRs (a)	SStrs (a)	StrRS (b)	SStrs (b)
10	0.945	0.891	0.143	0.899	0.228
20	0.664	0.964	0.187	0.798	0.133
30	0.557	0.801	0.230	0.745	0.100
40	0.485	0.719	0.241	0.661	0.070
50	0.430	0.664	0.244	0.556	0.059
60	0.392	0.536	0.209	0.469	0.048
70	0.361	0.402	0.177	0.351	0.043
80	0.338	0.203	0.152	0.092	0.096

Srs spatial random sampling, SStrs spatial stratified sampling model
a stratified by soil type, b stratified by geological surface

Table 2.3 Design effect of (spatial) stratified sampling by hierarchical cluster and geochronology

Sample sizes	Models			
	StrRs (c)	SStRs (c)	StrRS (d)	SStRs (d)
10	0.931	0.172	0.889	0.261
20	0.788	0.228	0.901	0.252
30	0.710	0.203	0.776	0.281
40	0.606	0.178	0.692	0.225
50	0.523	0.156	0.612	0.193
60	0.405	0.127	0.474	0.150
70	0.293	0.106	0.359	0.137
80	0.150	0.087	0.155	0.101

SStRs spatial stratified sampling model

c stratified by hierarchical cluster, *d* stratified by geochronology

knowledge layer may be very far apart and even spatially separated from the other layers. Spatial stratified sampling not only makes sure that variation is small in the same layer and large between layers, but also requires the layer is spatially continuous. In these four stratification methods, the second meet the requirement of keeping objects in the same layer distributed spatially continuous. So from Tables 2.2 and 2.3, it is seen that *b* has lower values for most sample sizes except at 10 and 80, and *c* has higher accuracy at samples, 10 and 80. Samples in the same knowledge layer in space have high spatial dependence, and $E\{C(X, Y)\}/n$ also increases. But at 10 and 80, the spatial dependence limitations become apparent. Hierarchical clustering stratifies samples by their Euclidean distance squared, so the properties of each layer are consistent within the sample, when compared to the results of other stratification methods, but samples are not together in space, resulting in high sampling efficiency outside the limitations of space.

2.4 Conclusions

In this research study, both spatial and non-spatial models have been investigated, with comparisons being made between five types of sampling models. Of these, sandwich spatial sampling and spatial stratified sampling were shown to have the highest efficiency, while the simple random model had the lowest efficiency. The efficiency of spatial random sampling and stratified sampling is shown to be dependent on the sample size. The stratified sampling model provides higher efficiency than spatial random sampling at both large and small sample sizes, but has lower efficiency at middle sizes or sizes between the two extremes. The sampling efficiency is also much dependent on the stratification method. As has been indicated, different methods can affect the sample distribution. For example, when stratified by different stratification method, even if two samples which are adjacent in space may be in different layers. The sample efficiency would be higher when the stratified factors can reflect the research objects' real spatial variance

better. Here, it means that the factor whose spatial distribution is closer to the real distribution of soil element content should be selected as knowledge stratification. At the same sample size. Spatial stratified sampling takes spatial dependence into account and keeps samples, together, in the same spacial knowledge layer, so stratification by geological surface has higher efficiency than other stratification methods. By comparing two kinds of stratification methods, by soil type and geochronology, it can be seen that, in the case of smaller sample sizes, stratification by soil type yields higher accuracy than by geochronology, but yields lower accuracy in larger sample sizes. Sandwich spatial sampling has the same efficiency as spatial stratified sampling for the sample mean and spatial mean variation. These values, however, can be calculated for each report layer. These report layers can then be divided arbitrarily, to meet the user's needs.

Acknowledgement The authors would like to thank Professor Chaoyang Wei and Tongbin Chen for their helpful discussions and constructive suggestions, and also appreciate the assistance of Luke Idrisk, Antoine Soullié and Hexiang Ba in the programming.

References

- Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York, pp 85–88
- Cressie N (1991) Statistics for spatial data. Wiley, New York
- de Zorzi P, Barbizzi S, Belli M, Mufato R, Sartori G, Stocchero G (2008) Soil sampling strategies: evaluation of different approaches. *Appl Radiat Isot* 66:1691–1694
- Chen DM, Wei H (2009) The effect of spatial autocorrelation and class proportion on the accuracy. *ISPRS J Photogramm Remote Sens* 64:140–150
- Wang GY, Pan M, Liu XD, Liang HH (1992) On the relationship between the concentrations of elements in soil and the types of soil-forming parental material in Shandong province, China. *Acta Sci Nat Univ Pekin* 28(4):475–485
- Haining RP (1988) Estimating spatial means with an application to remote sensing data. *Commun Stat Theory Methodol* 17(2):537–597
- Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge
- Hedger RD, Atkinson PM, Malthus TJ (2001) Optimizing sampling strategies for estimating mean water quality in lakes using geostatistical techniques with remote sensing. *Lakes and Reservoirs: Research and Management* 6(4):279–288
- Wang JF, Liu JY, Zhuang DF, Li LF, Ge Y (2002) Spatial sampling design for monitoring the area of cultivated land. *Int J Remote Sens* 23(2):263–284
- Wang JF, Haining RP, Cao ZD (2010) Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. *Int J Geogr Inf Sci* 24(4):523–543 doi:10.1080/13658810902873512
- Wang JF, Jiang CS, Li LF, Hu MG (2009) Spatial sampling and statistical inference (in Chinese). Science Press, Beijing, pp 17–20
- Li LF, Wang JF, Liu JY (2005) Optimal decision-making model of spatial sampling for survey of Chinas land with remotely sensed data. *Sci China Ser D* 48(6):752–764
- Zhang ZD, Zhang KL, Wu JL, Zhang T, Zheng XY (2008) Spatial information analysis of characteristics of soil trace element contents in high incidence areas of birth defects. *J Zhejiang Univ (Agric Life Sci)* 34(6):684–691

Advances in Spatial Data Handling and GIS

14th International Symposium on Spatial Data Handling

Yeh, A.G.O.; Shi, W.; Leung, Y.; Zhou, C. (Eds.)

2012, XXII, 242 p., Hardcover

ISBN: 978-3-642-25925-8